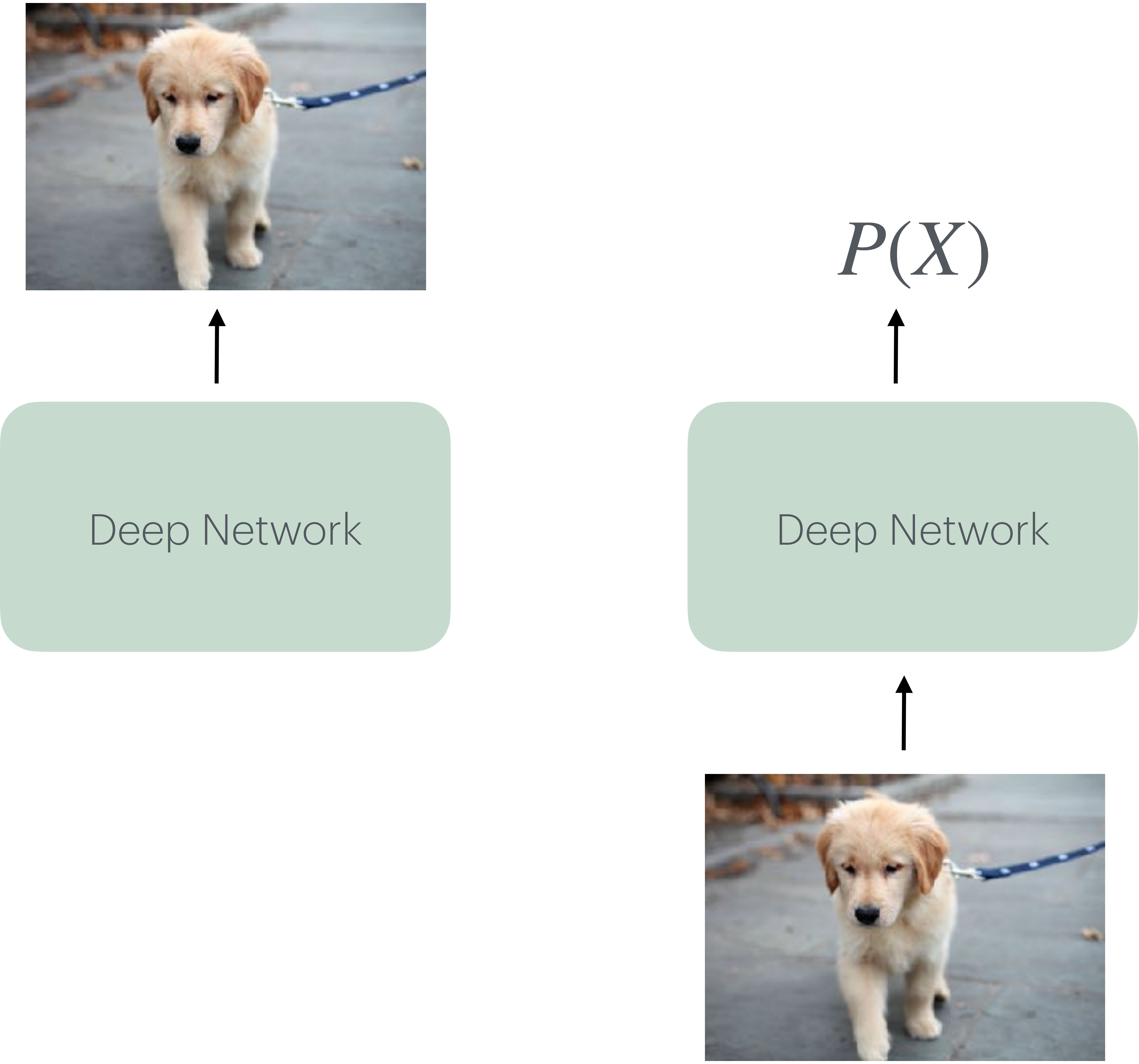
# Vector Quantization

Philipp Krähenbühl, UT Austin

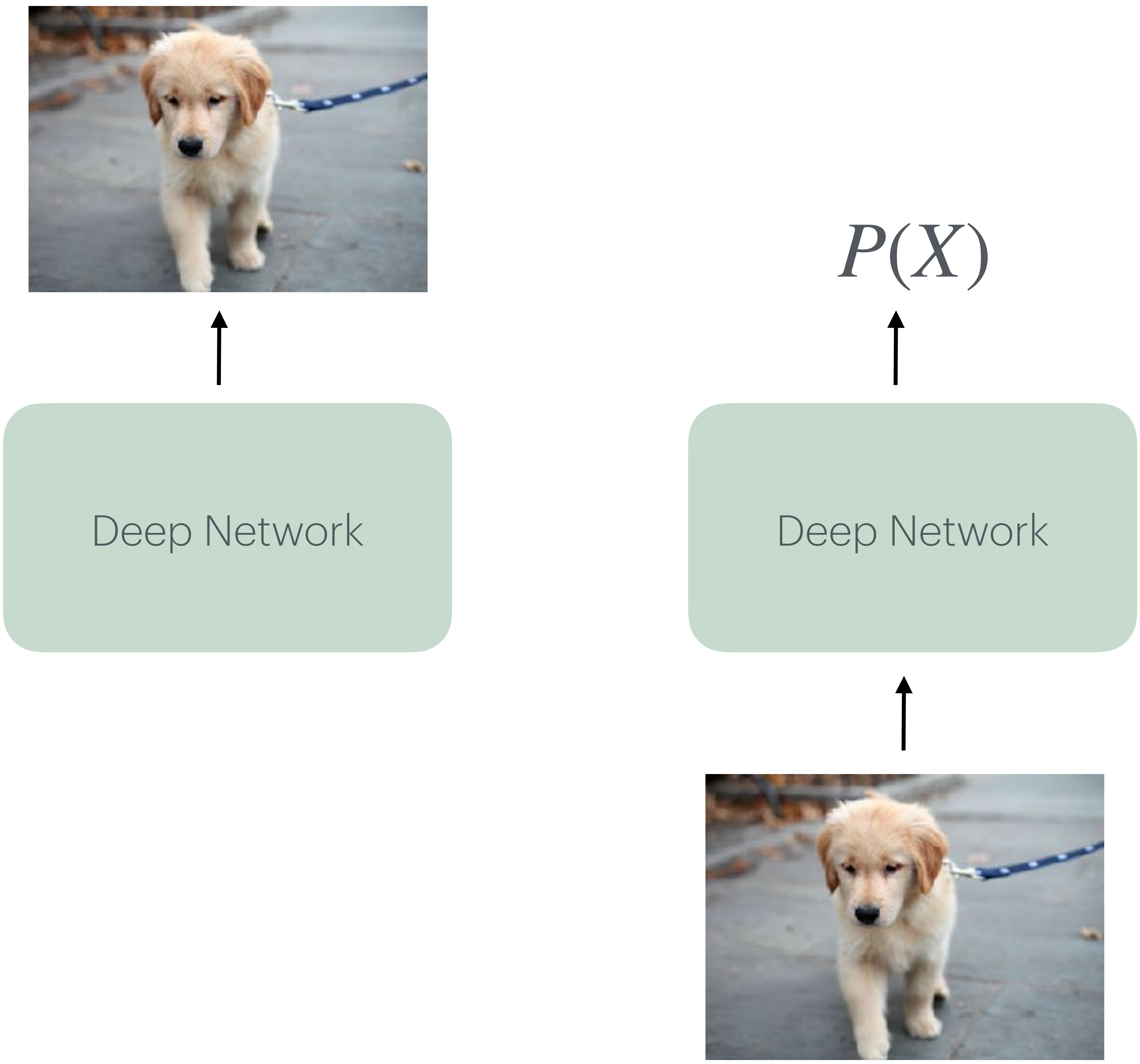
# Generative models

- Two tasks of a generative model $P(X)$
  - Sampling: $x \sim P(X)$
  - Density estimation: $P(X = x)$



# Generative modeling is hard

- Density estimation $P(X = x)$
  - How to ensure $\sum_x P(x) = 1$ for all $x$
  - Impossible to compute (in general)
- Sampling $x \sim P(X)$
  - What is the input to the network?



# Generative models

## Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x|z)$ or $f: z \to x$

$z$ Deep Network

Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

Deep Network $P(X)$

# Auto-regressive models

## Issues


$$P(x) = P(x_1)P(x_2 | x_1)P(x_3 | x_1, x_2)P(x_4 | x_1 \ldots x_3)\ldots$$

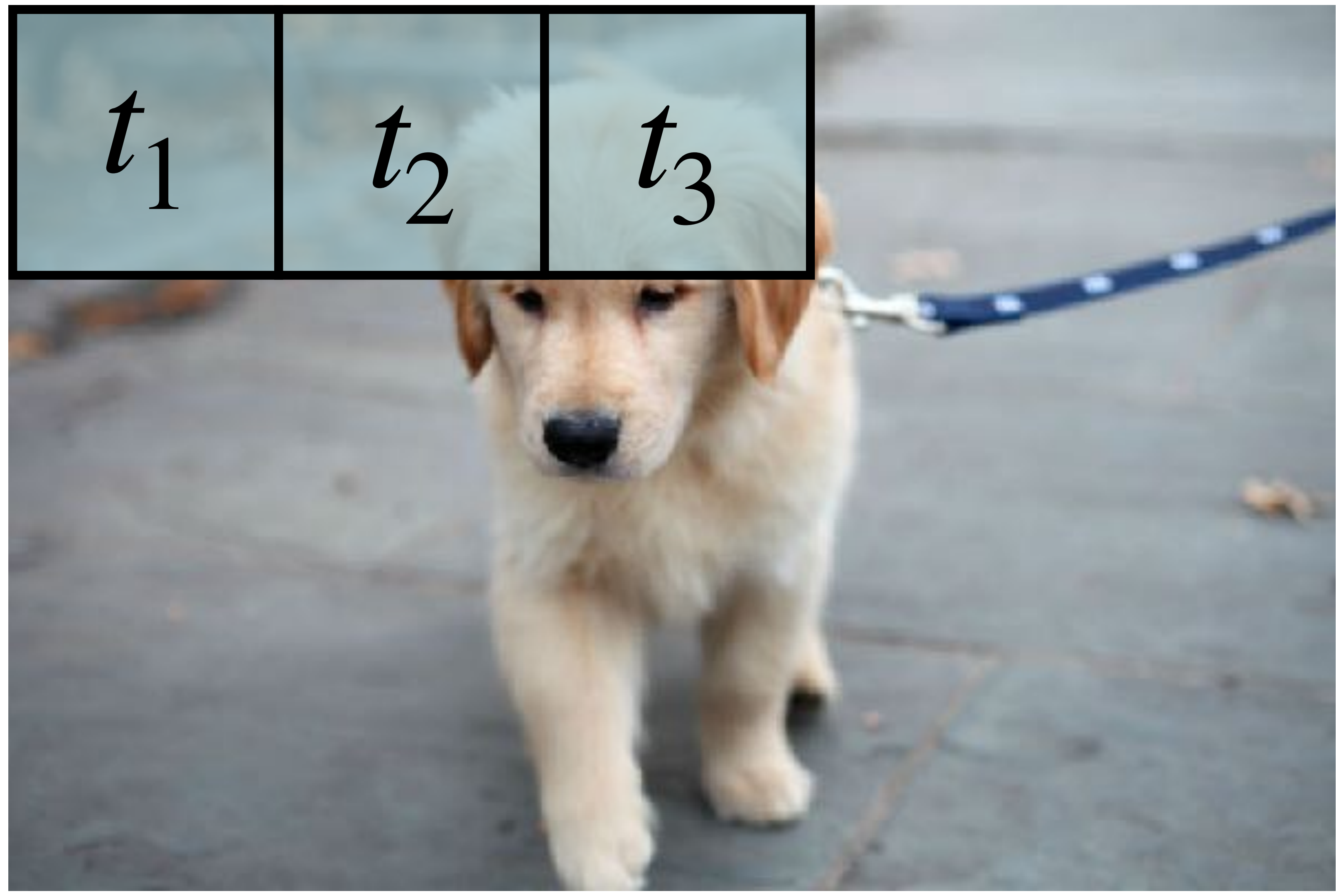
- Difficult learning problem for long sequences (requires good model)

[1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
[2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022

# Tokenization

- Image [1]
  - Convert patch $p_i$ of pixels into token $t_i \in \{1, \ldots, K\}$
- Text [2]
  - Convert set of characters into token
- Protein-sequence [3]
  - Convert local protein structure to token


Vanilla auto-regressive model




Tokenized auto-regressive model

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017
[2] Language models are unsupervised multitask learners. Alec Radford, et al. 2019
[3] Simulating 500 million years of evolution with a language model. Thomas Hayes, et al. 2024

# Auto-regressive models on tokens

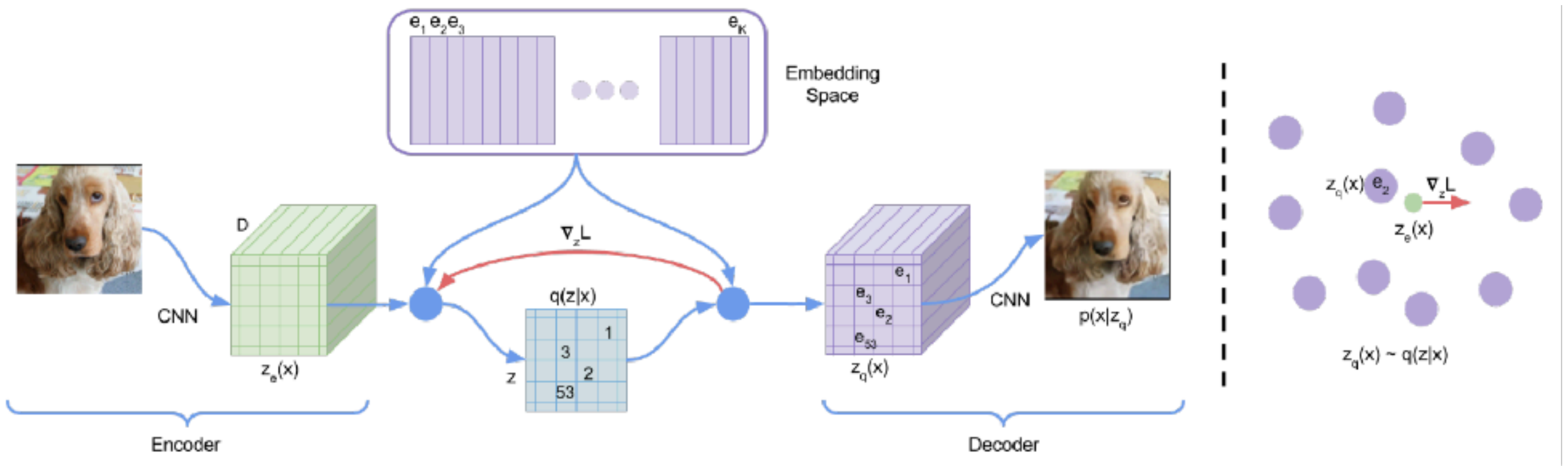
$$P(\mathbf{t}) = P(t_1)P(t_2 \mid t_1)P(t_3 \mid t_1, t_2)P(t_4 \mid t_1 \ldots t_3)\ldots$$

- Shorter sequence = easier to learn structure

[1] MAGVIT: Masked Generative Video Transformer. Lijun Yu, et al. 2023

# Learning Tokenization

## Vector Quantization



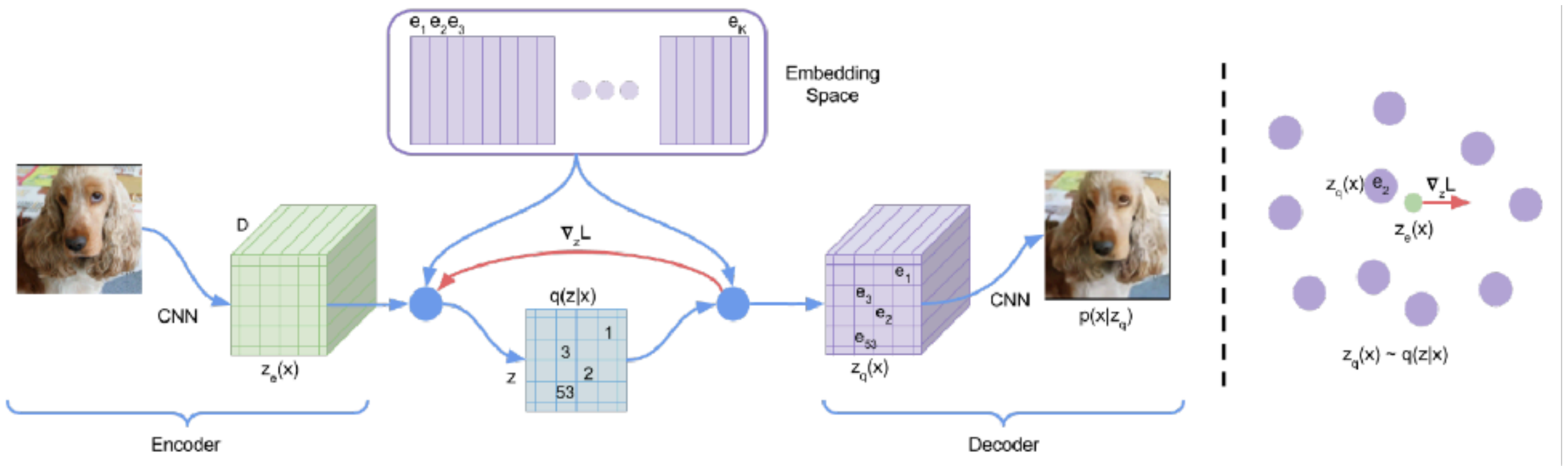

- Input: Image (or patch)
  $x \in \mathbb{R}^{H \times W \times 3}$

- Output: "Image" of tokens
  $z \in \{1 \ldots K\}^{h \times w}$

- Why is this hard to learn?
  - $z \to x$ (easy, reconstruction)
  - $x \to z \to x$ (hard, $z$ is discrete and non-differentiable)

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

# VQ-VAE

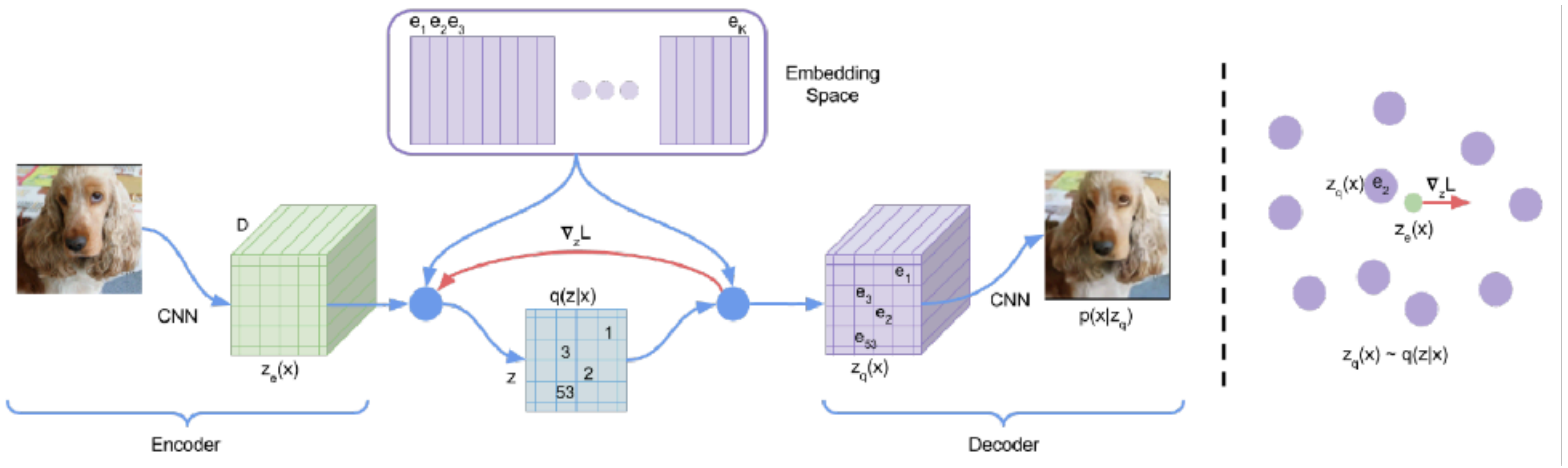



- Variational Auto-Encoder
  - Decoder $P_D(x|z)$ Encoder $Q(z|x)$
- Vector Quantizer
  - $q(z) = \arg\min_{e_k} ||z - e_k||$
  - Learn codebook $\{e_1 \ldots e_K\}$
  - What is $\nabla q(z)$?

[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

# VQ-VAE

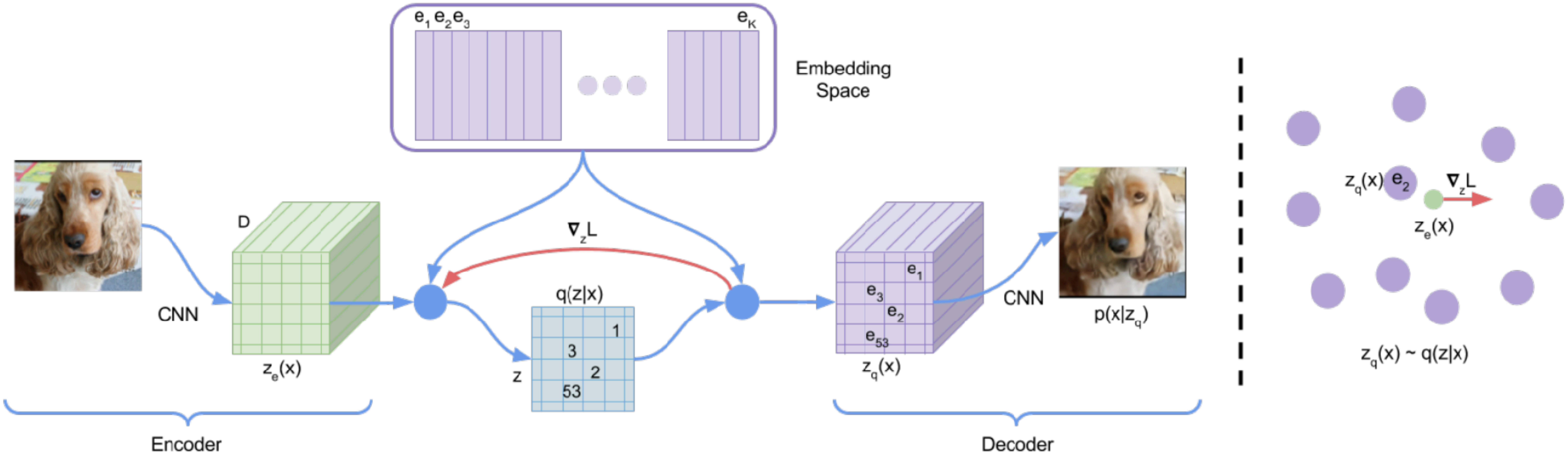
## Gradient




- What is $\nabla q(z)$?
  - Let's assume $\nabla q(z) = \mathrm{I}$ (identity)
  - Straight-Through Estimator
    - Works in practice because errors average out over large enough batches
    - No reason it should work

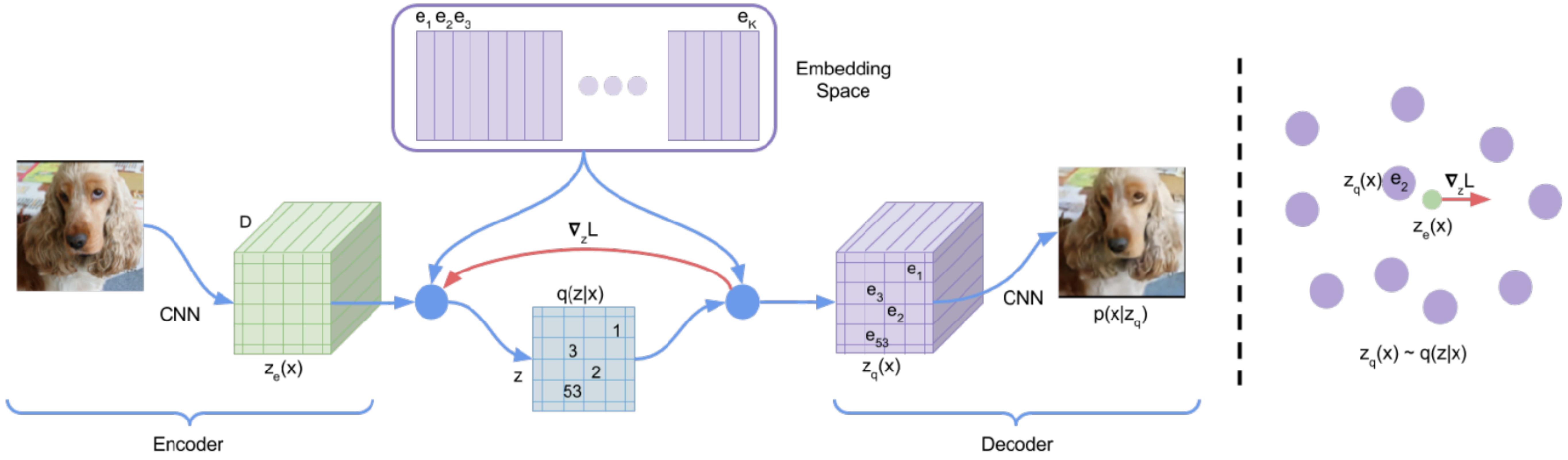
[1] Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. Yoshua Bengio, et al. 2013

# VQ-VAE




[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

# VQ-VAE




- Only as good as VAE
- Does not scale well with codebook size
  - Codebook grows exponentially in #bits
  - Many entries → sparse gradients
  - Slow

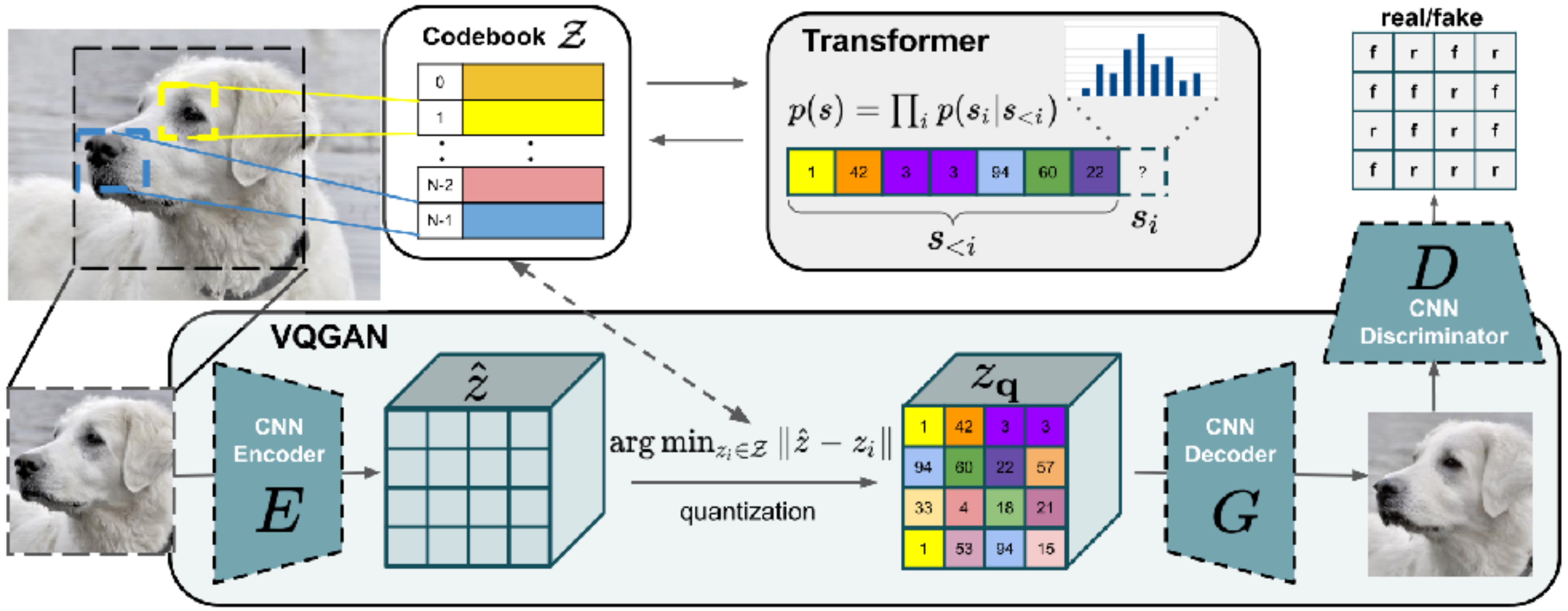
[1] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017

# VQ-GAN

- Replace VAE with GAN
    - Auto-encoder with vector quantization

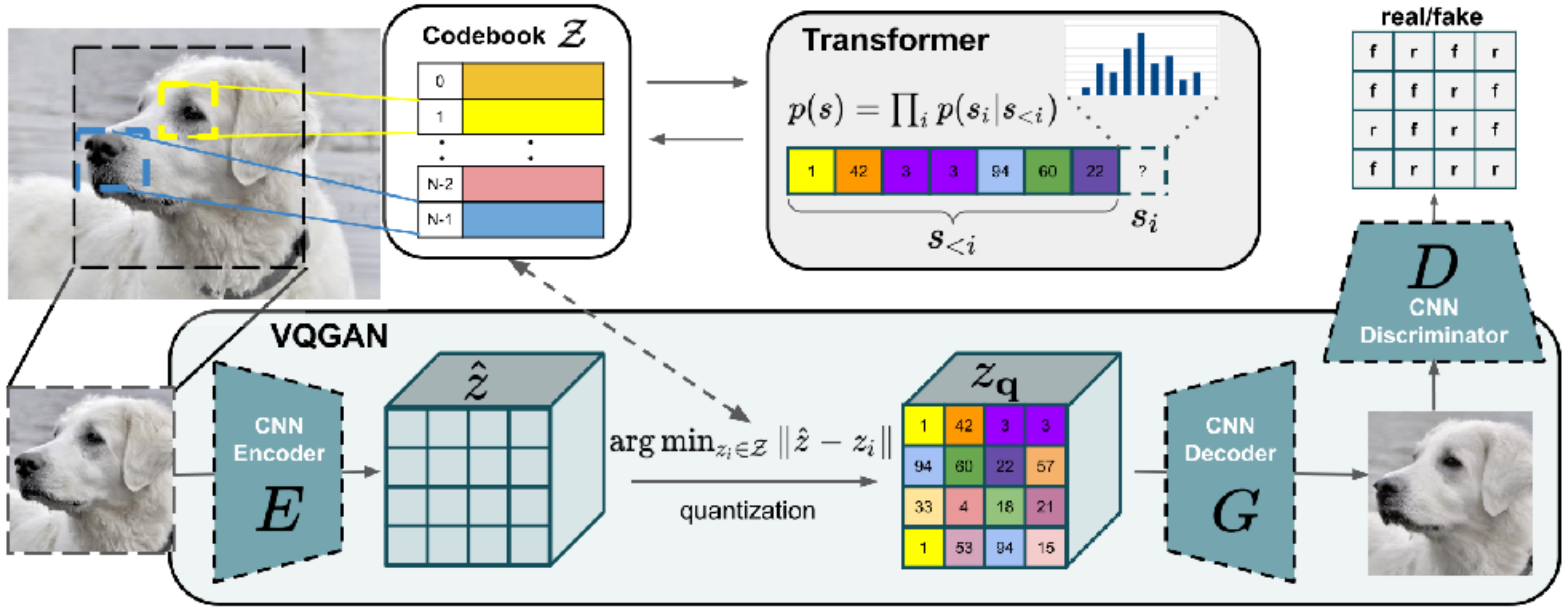
      $$q(z) = \arg\min_{e_k} \|z - e_k\|$$

    - GAN + Reconstruction loss
- Learn a sequence model on top
- Default image tokenizer nowadays




[1] Taming transformers for high-resolution image synthesis. Patrick Esser et al. 2021

# VQ-GAN



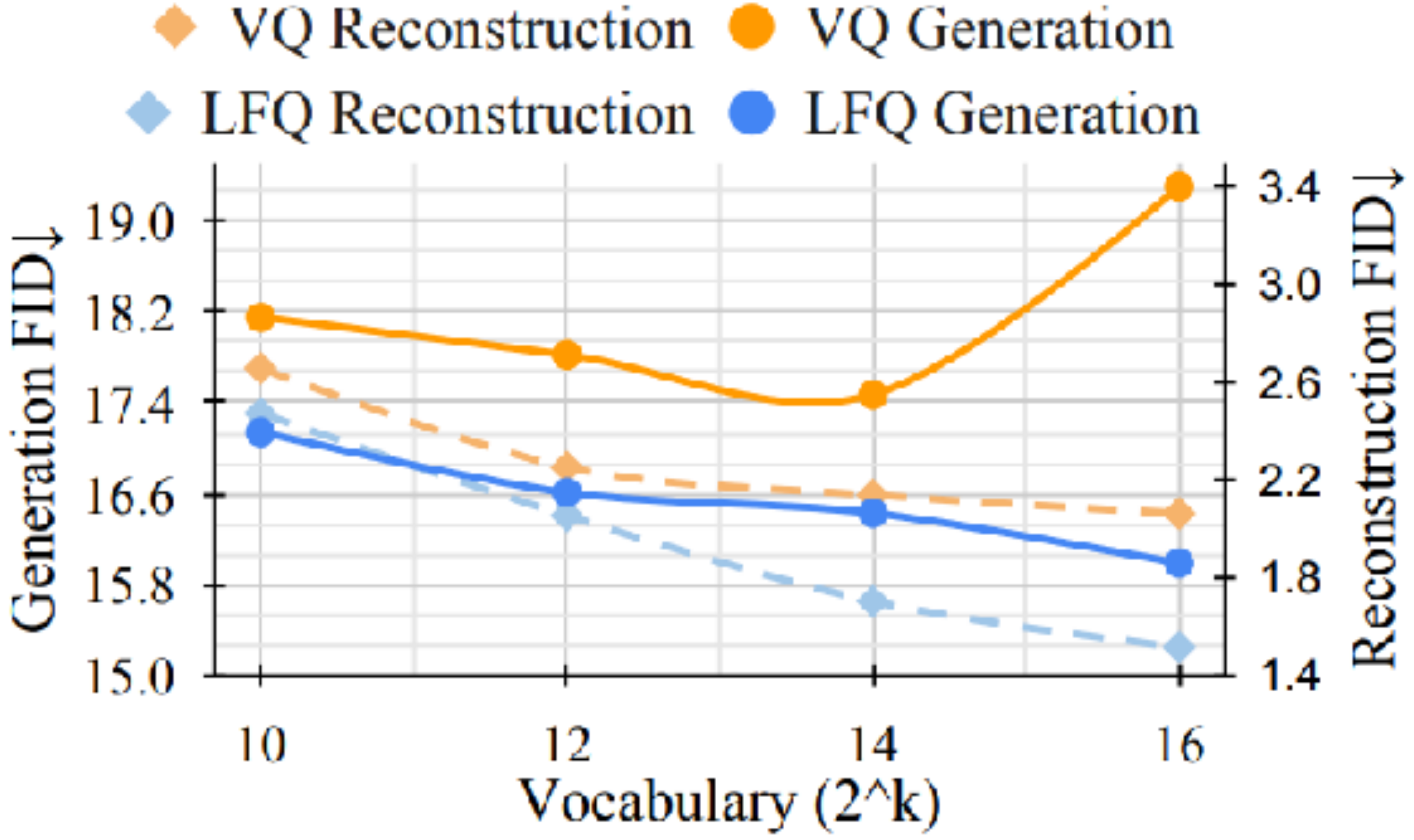

- Great tokenizer, ok sequence model
- Does not scale well with codebook size
  - Codebook grows exponentially in #bits
  - Many entries → sparse gradients
  - Slow

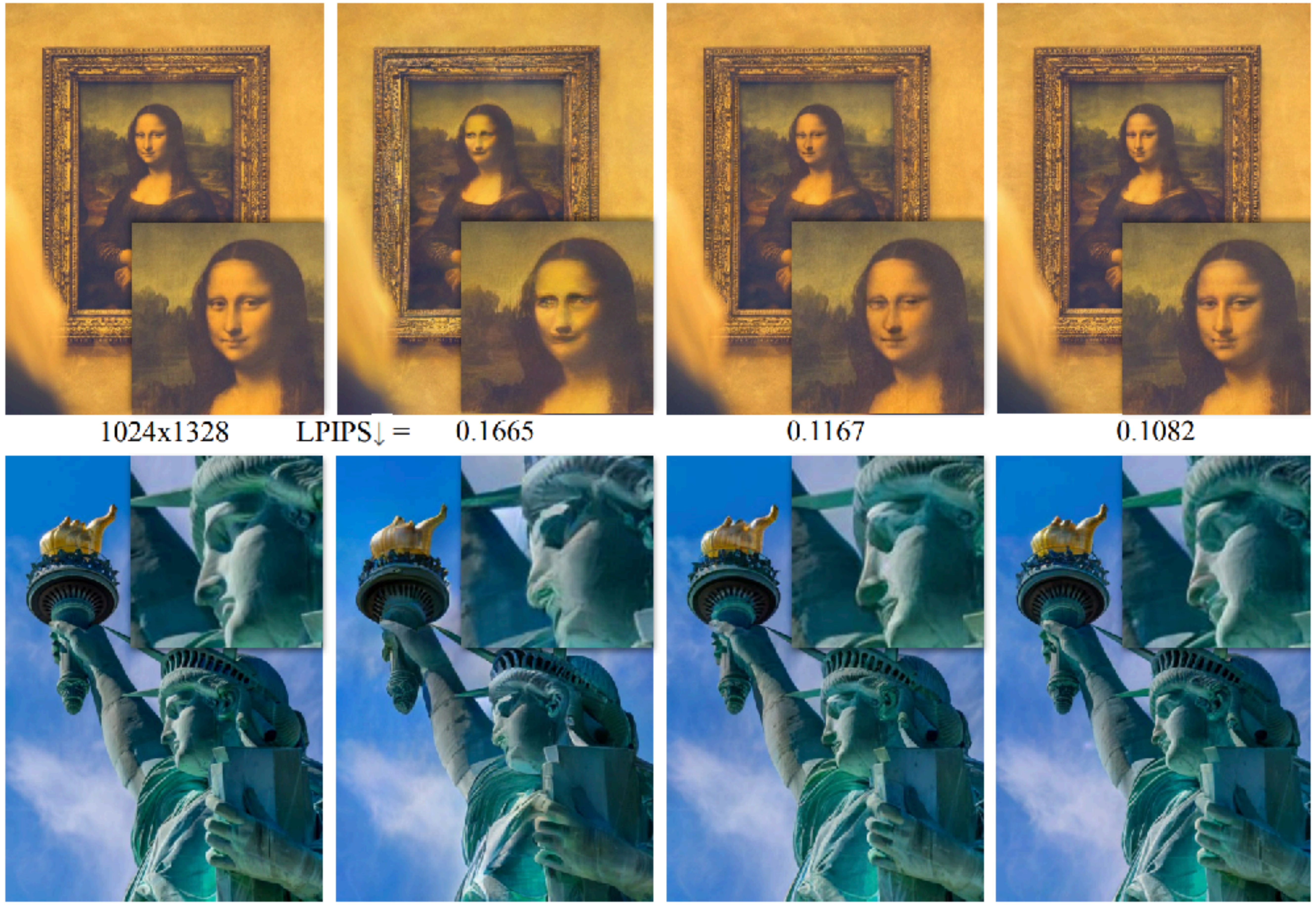
[1] Taming transformers for high-resolution image synthesis. Patrick Esser et al. 2021

# LFQ

## Lookup-Free Quantization

- Different quantizer
  - $q(z) = \mathrm{sign}(z)$ where $\mathrm{sign}(z_i) = 1_{[z_i \le 0]} - 1_{[z_i > 0]}$
- Scales linearly with #bits in bottleneck
- No learned parameters

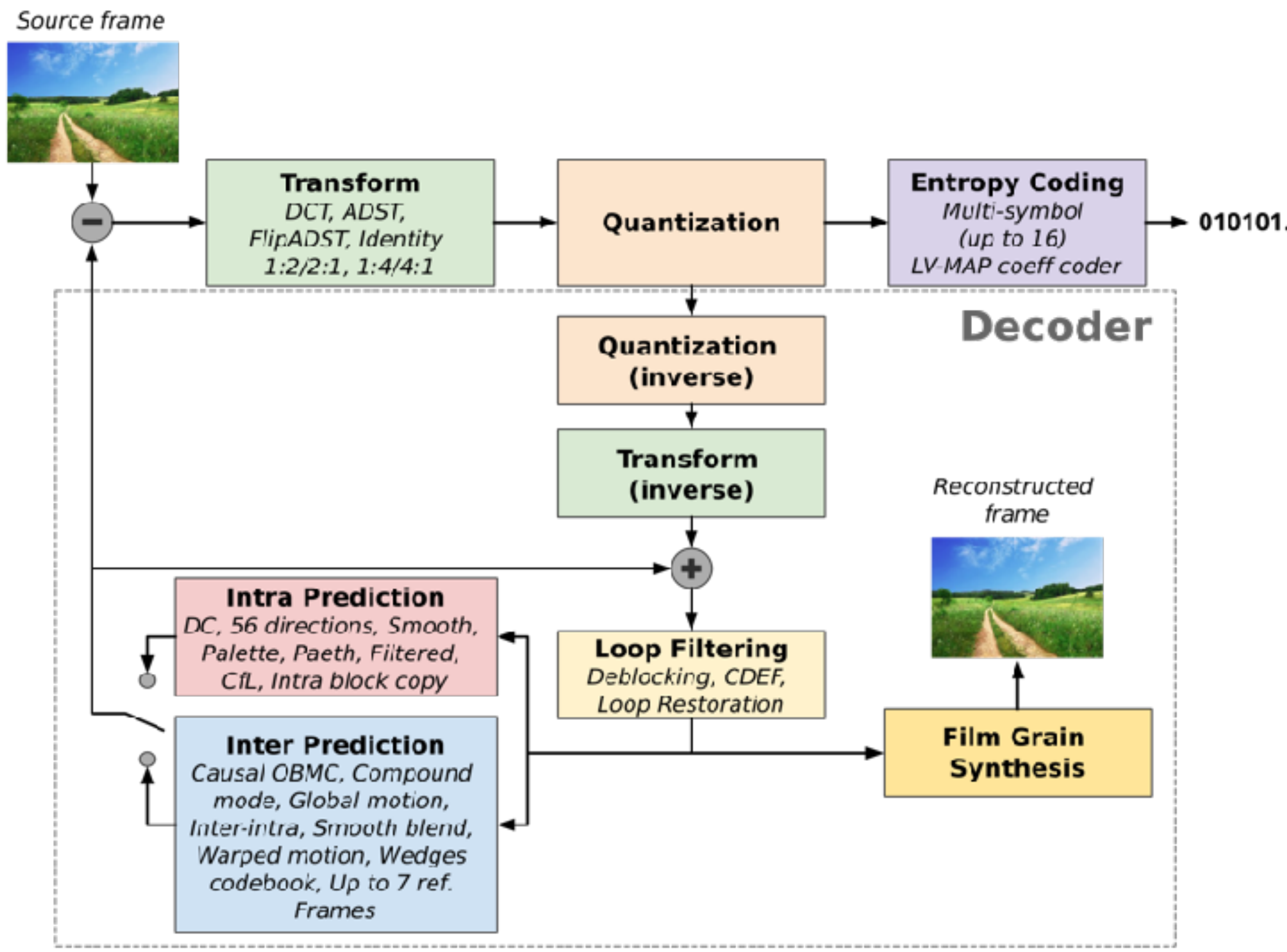



[1] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation. Lijun Yu, et al. 2024

# Generation vs Compression

- Auto-regressive model
  - Lossless compression (fancy gzip)
- Tokenization (VQ)
  - Lossy compression
- Similar to how JPEG most video codecs work



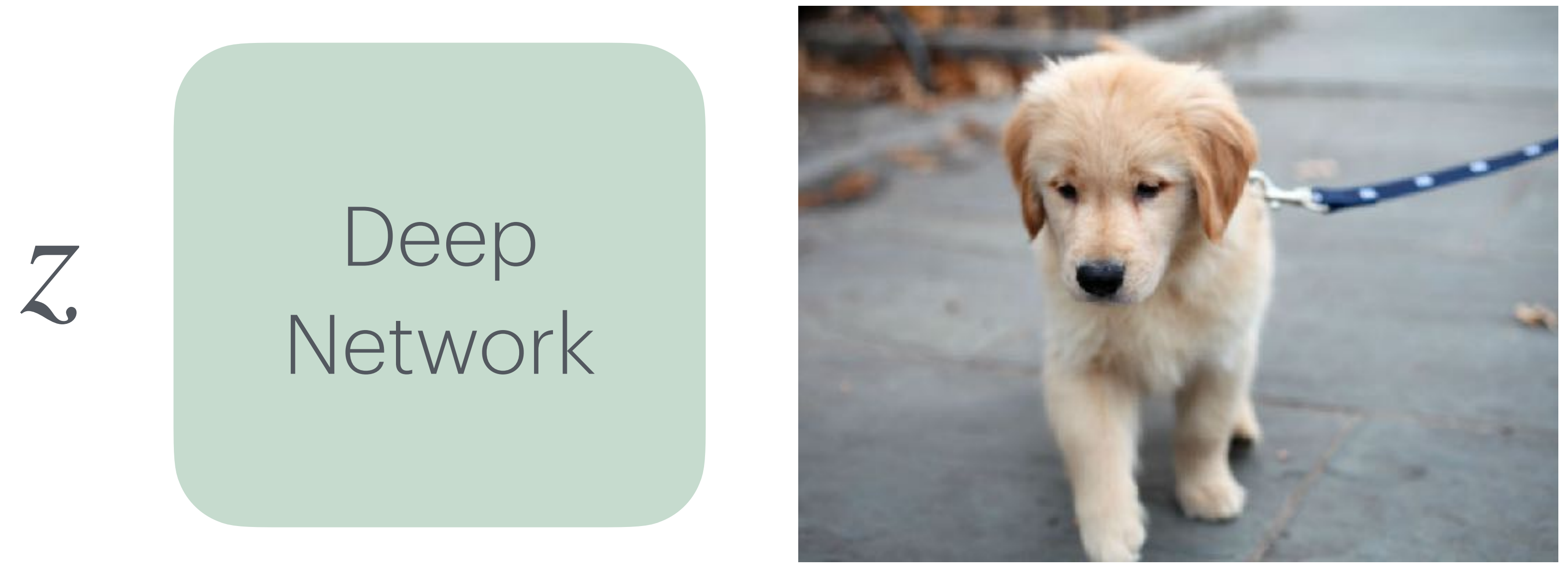

Source: https://commons.wikimedia.org/wiki/File:The_Technology_Inside_Av1.svg

# Generative models

## Two kinds of models

Sampling based $x \sim P(X)$

- Sample $z \sim P(Z)$
- Learn transformation
  - $P(x \mid z)$ or $f : z \rightarrow x$




Density estimation based $P(X)$

- Learn special form of $P(X)$
- Model specific sampling / generation

# References


- [1] WaveNet: A Generative Model for Raw Audio. Aaron van den Oord, et al. 2016
- [2] Long Video Generation with Time-Agnostic VQGAN and Time-Sensitive Transformer. Songwei Ge, et al. 2022
- [3] Neural Discrete Representation Learning. Aaron van den Oord, et al. 2017
- [4] Language models are unsupervised multitask learners. Alec Radford, et al. 2019
- [5] Simulating 500 million years of evolution with a language model. Thomas Hayes, et al. 2024
- [6] MAGVIT: Masked Generative Video Transformer. Lijun Yu, et al. 2023
- [7] Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. Yoshua Bengio, et al. 2013
- [8] Taming transformers for high-resolution image synthesis. Patrick Esser et al. 2021
- [9] Language Model Beats Diffusion -- Tokenizer is Key to Visual Generation. Lijun Yu, et al. 2024